自然言語処理 110－13
（1995．11．18）

非対訳コーパスを用いた訳語関係の抽出

田中久美子
東京大学工学部

岩崎英哉
東京大学教育用計算機センター

対訳でない独立のコーパスを用いて訳語関係を抽出する手法を提案する。「第一言語で共起する二つの語の訳語は第二言語でも共起する」という仮定に基づき、第一言語の共起情報を第二言語に翻訳することを確率行列の枠組で形式化した。その上で、第一言語の共起情報の翻訳と、第二言語の元の共起情報が類似する様に翻訳行列を最適化する。最適である時は訳語関係の曖昧性が文脈に基づいて最も解消されている場合となる。まず、局所文脈に基づいた訳語の選択を行えることを同枠組で検証した。さらに、第一、二言語で非対訳コーパスから大域的に訳語関係を抽出するために、翻訳行列を逐次的に最適化する原理を示し、有効性を示す定性的な小実験を行なった。

Extraction of Word Translations from Non-Alinged Corpus

Kumiko TANAKA
The University of Tokyo
Faculty of Engineering

Hideya IWASAKI
The University of Tokyo
Educational Computer Centre

A method to extract word translations from non-aligned corpus is proposed. The assumption "translations of two co-occurring words in the source language also co-occur in the target" is adopted and is represented in the stochastic matrix formulation. The best translation matrix gives the co-occurring information translated from the source, which is close to the co-occurring information in the target, when the ambiguity of translational relation is resolved according to the non-alinged corpus. The algorithm to obtain the best translation matrix is introduced. Small experiences are performed to evaluate the effectiveness of ambiguity resolution and to obtain the best translation matrix.

1 研究の目的意義

自動翻訳の補助および語の訳語関係の抽出を目的として 対訳コーパスの対応付けの研究がさかんに行われている。訳語関係の抽出に焦点を当てると、文対応など複数語の単位で対応付けを行った後、それをもとに最尤法などで語対応を付けて初めて頻出する訳語対応を取りだすことができる。

対訳コーパスに最尤法などで語対応を付けるにはトレーニングのために大量の対訳コーパスが必要である [Bro93]。しかし、大量の対訳コーパスが存在するような二言語は常に一方が国際性の高い言語に限定され、どの言語も国際性の高い言語との間にはそれなりの品質の辞書が存在する。語彙情報の抽出が急務であるのは、むしろ大量の対訳コーパスが入手しにくい特殊分野や特殊な言語間ではなかろうか。この場合、対訳コーパスが存在しなくとも単言語ではそれぞれ独立にコーパスは存在するので、それを用いて訳語関係の抽出を行なうことができれば理想的である。

特殊分野や特殊な言語間の辞書を編纂するには、元言語と国際語、国際語と目的言語の二冊の辞書を変形して訳語候補を得て辞書の原型とすることがさかんに行われている。この時には、中間の言語における語義の曖昧性から生じる不適当な訳語候補をとり除くことが必要である。すなわち、正誤含んで訳語候補がすでに存在する時に正しいものを抽出する方法を確立する必要があり、これは古来よりさかんに行われてきた文脈に応じて適当な訳語を選び出す研究と関係深い。

本稿では、正誤含めていくつかの訳語候補がすでに存在する時に、対訳でない二言語で独立のコーパスを用いて局所あるいは大域文脈で適当な訳語関係を選び出す手法を示す。その要となるのは「第一言語で共起する二つの語の訳語は第二言語でも共起する」という仮定である。この仮定に基づいた二言語の訳語関係の曖昧性の解消法を第二章で提示する。第三章では曖昧性が実際に解消されることを、局所文脈に基づいた訳語の選択を行なうことを通して論ずる。第四章では二つの独立なコーパスから訳語関係を大域的に抽出する手法を示す。第五章では実験を行なって有効性を検証する。

以下では第一言語を L_A、第二言語を L_B と記す。また、実験では第一言語を英語、第二言語を日本語とする。例の引用には**日本語**や English といった字体を用いる。

2 仮定と曖昧性の解消法

2.1 仮定とその形式化

「第一言語で共起する[1]二語の訳語は、第二言語でも共起する」という仮定を用いる。たとえば英語で doctor と nurse は共起し、日本語でもこれらに対応する訳語の**医者**と**看護婦**は共起する。

Rapp[Rap95] は英語とドイツ語の間でこの仮定が成り立つことを検証した。L_A の語 a_i と a_j の共起度を第 (i,j) 成分とする行列 A と L_B の語 b_k と b_l の共起度を第 (k,l) 成分とする行列 B があって、しかも A と B の a_i と b_i が一対一対応する時には A と B が類似するということの検証を行なった。

しかし、語の対応は通常は一対多であるから、本稿では以下のように形式化する。まず A の第 (i,j) 成分を a_i と a_j の共起確率[2]とする。すると A は全成分の和が 1.0 の対称行列となる。B も L_B 上で同様に定義する。さらに L_A から L_B への翻訳行列 T を a_i から b_k への翻訳確率 $p(b_k|a_i)$ を第 (i,k) 成分とする行列として定義する。T は正方行列とは限らない。また T の行和は 1.0 である。

この時、L_A の共起情報は L_B に翻訳することができ (図 1参照)、b_k と b_l の共起確率は、

$$\sum_{u,v} p(b_k|a_u)p(a_u,a_v)p(b_l|a_v) \qquad (1)$$

となる。行列表現で書けば

$$T^t A T \qquad (2)$$

であり、これは全成分の和が 1.0 の対称行列となる。仮定によればこれが B と類似する。

2.2 曖昧性の解消

前節の例を再考すると、doctor の訳語は**医者**であって**博士**ではないということも分かる。それは、doctor と共起する語が nurse であり、nurse に対応する**看護婦**と共起する日本語は**博士**よりもむしろ**医者**であるからである。すなわち、二言語の語の共起関係の類似性を用いると、訳語関係の曖昧性の解消を行なうことができる。

すなわち、前節の行列の枠組みで記述すると、A や B が既知の時に $|X-Y|$ を行列 X と Y の適

[1] 二つの語が共起するとは、その二つの語が文章の中で至近距離で生起することを指す。

[2] 共起確率は a_i と a_j が共起した頻度を A 中の語の共起頻度の総和で割ったもので近似する。

$$\begin{array}{ccc} & T & \\ a_u & \longrightarrow & b_k \\ A\updownarrow & & \Updownarrow T^tAT \text{ vs. B} \\ a_v & \longrightarrow & b_l \\ & T & \end{array}$$

図 1: 確率行列計算T^tATの図式

当な距離として、

$$min|T^tAT - B| \qquad (3)$$

となる様なTを求めることが訳語関係の曖昧性の解消であるということとなる。$|X-Y|$に関しては成分の二乗和、最大の成分の絶対値の最大値などさまざまな距離が考えられるであろう。

2.3 独立性に関する考察

式 (1) は$p(b_k|a_u)$と$p(b_l|a_v)$の積をとっているが、そこでは、

$$p(b_k, b_l|a_u, a_v) \simeq p(b_k|a_u)p(b_l|a_v) \qquad (4)$$

という近似が行なわれている。

$$p(b_k, b_l) = \sum_{u,v} p(b_k, b_l|a_u, a_v)p(a_u, a_v) \qquad (5)$$

は常に成り立つが、これと式 (1) が類似するという前節の議論を比較すると式 (4) が導出される。

$p(b_k, b_l|a_u, a_v)$を求めることが式 (5) に最も忠実である。しかし、実際には$p(b_k, b_l|a_u, a_v)$は、単語と単語の一対一の対応の確率を表現していないので、この確率をどのように利用して曖昧性を解消するかが大きな問題となる。また、$p(b_k, b_l|a_u, a_v)$を求めるには一辺が一言語の語の数の2乗という莫大な大きさの行列計算が必要であり、計算量の面から考えても実際的ではない。

式 (4) は、

$$\frac{p(a_u, a_v)}{p(a_u)p(a_v)} \simeq \frac{p(a_u, a_v, b_k, b_l)}{p(a_u, b_k)p(a_v, b_l)}$$

と等価である。仮定のL_AとL_Bにおける対称性を考慮すると、

$$p(a_u, a_v|b_k, b_l) \simeq p(a_u|b_k)p(a_v|b_l)$$

という近似も導入する必要があり、

$$\frac{p(a_u, a_v)}{p(a_u)p(a_v)} \simeq \frac{p(a_u, a_v, b_k, b_l)}{p(a_u, b_k)p(a_v, b_l)} \simeq \frac{p(b_k, b_l)}{p(b_k)p(b_l)} \qquad (6)$$

が成り立つ。$p(x,y)/p(x)p(y)$がxとyの独立性の度合を測るのに有効であることが Church ら [Chu89] によって示されたことをふまえると、式 (4) の近似はL_AとL_Bで類似する共起度を示す二語を求めることを表現しており、これはまさに仮定そのものである。すなわち、式 (4) が成り立つような$p(b_k|a_u)$を求めることは仮定が成り立つような$p(b_k|a_u)$を求めることを表す。

3 局所文脈による曖昧性の解消

本節では式 (3) によって曖昧性が実際に解消される仕組みを示す。従来より局所文脈に基づいて訳語を自動的に選択する手法が議論されてきたが、本章ではその問題を取りあげる。局所文脈を扱うには、Aを局所的に作成して式 (3) を計算する。

3.1 行列間の距離

二つの行列間距離$|X-Y|$を以下のように二種類定義する。

$|X-Y|_{pat}$は、Rapp[Rap95] も用いているものであるが、XもYも成分は0以上であることをふまえて、行列XとYの成分の符号によって距離を測る。各成分ごとに共に正あるいは零であれば0、一方のみ正であれば距離を1とし、全成分の距離の和をとる。たとえば、

$$\left|\begin{pmatrix} 2 & 0 \\ 0 & 3 \end{pmatrix} - \begin{pmatrix} 4 & 1 \\ 0 & 0 \end{pmatrix}\right|_{pat}$$

は$(0,0)$成分の距離は0, $(0,1)$成分の距離は0, $(1,0)$成分の距離は1, $(1,1)$成分の距離は1で、合計2となる。

$|X-Y|_{squ}$は行列の成分の差の二乗和である。上の例では、

$$(2-4)^2 + (0-1)^2 + (0-0)^2 + (3-1)^2 = 9$$

と計算される。

3.2 doctor の例

doctor に対する訳語候補として**医者**と**博士**の二語があり、doctor は「The doctor nursed the patient.」という文脈で起こったとする。文脈によると適当な訳語は明らかに**医者**の方である。

この文章に基づいてAを次の様に作成する。

	doc	nur	pat
doctor	0	1/6	1/6
nurse	1/6	0	1/6
patient	1/6	1/6	0

また、T は以下のようなものであったとする。

	医者	看護する	患者	博士	大学
doc	t_{11}	0.0	0.0	t_{41}	0.0
nur	0.0	1.0	0.0	0.0	0.0
pat	0.0	0.0	1.0	0.0	0.0

大学は**博士**の主な共起語として B にあるので、T に導入した。また、実際には nurse や patient も多義であるがここでは簡単のため上のような行列とした。

B はコーパス全体から共起性の統計をとったもので、関連する語にのみ注目し、たとえば以下のようになったとする。

	医	看	患	博	大
医	0.0	0.1	0.1	0.0	0.0
看	0.1	0.1	0.1	0.0	0.0
患	0.1	0.1	0.0	0.0	0.0
博	0.0	0.0	0.0	0.05	0.1
大	0.0	0.0	0.0	0.1	0.05

医者と**博士**のどちらがより適当な訳語を知りたいので、T の t_{11}, t_{41} の一方の成分を 1.0 もう一方を 0.0 と試験的に置いて T^tAT を計算する。$t_{11} = 1.0$, $t_{41} = 0.0$ と置いて、つまり doctor と**医者**が対応するとして計算すると、T^tAT は

	医	看	患	博	大
医	0	1/6	1/6	0	0
看	1/6	0	1/6	0	0
患	1/6	1/6	0	0	0
博	0	0	0	0	0
大	0	0	0	0	0

となる。この時、

$$|T^tAT - B|_{pat} = 5.0$$
$$|T^tAT - B|_{squ} = 0.0734$$

となる。次に $t_{41} = 1.0$, $t_{11} = 0.0$ と置いて、つまり doctor と**博士**が対応するとして T^tAT 計算すると、

	医	看	患	博	大
医	0	0	0	0	0
看	0	0	1/6	1/6	0
患	1/3	0	1/6	1/6	0
博	0	0	0	0	0
大	0	0	0	0	0

となる。この時、

$$|T^tAT - B|_{pat} = 12.0$$
$$|T^tAT - B|_{squ} = 0.1405$$

である。この二つの結果から、doctor と**医者**が対応するとした時の方が doctor と**博士**が対応するとした時よりも距離は小さい。すなわち、**医者**の方がより適当な訳語ということになる。

この様に訳語候補がすでにいくつか存在して、そのうちのどれが適当な訳語かを判断するには、各訳語候補につきそれが正しいと仮定して式 (3) を試験的に計算し、距離の最も小さいものを選ぶことで、局所的な文脈に依存した訳語関係を判断できる。

3.3 過去の研究との関連

Dagan [Dag94] [Dag91] は、L_A で曖昧性を解消してから L_B の訳語を選択するのではなく、訳語のすべての可能性を L_B で挙げ、L_B における共起関係を用いて訳語を選択しようとした。訳語を選択したい語を a_u とし、文構造をふまえた上で重要な共起語を選び (たとえば述語に対して目的語など)、それを a_v とする。a_u, a_v に対するあらゆる翻訳の可能性 (b_i,b_j と記述する) を二言語辞書で抽出し、L_B での共起頻度 ($freq$ と記述する) を測る。それを用いて、

$$p(b_k, b_l|a_u, a_v) = \frac{freq(b_k, b_l)}{\sum_{i,j} freq(b_i, b_j)}$$

と最尤推定する。最も $p(b_k, b_l|a_u, a_v)$ が大きいものを検定した上で、b_k を a_u の訳語とする。本手法との違いは、Dagan が $p(b_k, b_l|a_u, a_v)$ を推定するのに対し、本手法は $p(b_k|a_u)$ を計算する枠組にのっとっている点である。したがって、Dagan の方法は a_u, a_v が局所文脈ですでに確定している場合にのみ曖昧性の解消として適用可能な手法である。それに対して本手法は次章で述べるような大域的な文脈における訳語対応を抽出するのにも適用可能である。

語をノード、共起や訳といった語の関係を枝とすると、語の共起関係や訳語対応はグラフとみなすことができる。AT の (i,k) 成分が正であることは、a_i からある共起語 a_u を経由して b_k までたどりつけることである。さらに、T^tAT も B も正となる成分は L_A の共起する二つの語に対して L_B で共起する二つの語に訳語があるということであり、グラフでは閉じた回路ができている時である。

この回路の概念は、電子辞書からの概念辞書抽出を二言語辞書からの回路を抽出することによって行なうという Tokunaga[Tok90] の枠組と類似している。しかし、そこでは共起関係は用いられておらず訳語対応に回路があればそれを概念辞書

の抽出に役立てることができないかという試みとなっている。

4 二言語間語彙情報の抽出

二言語で独立のコーパスを用いて訳語関係の抽出を行なう。前章では A は局所的な文章から作成した。それに対しここでは A を大域的な共起確率から作成して式 (3) が最小となるような T の求め方を論ずる。尚、この章では距離は二乗和の $|A-B|_{squ}$ のみ扱う。

4.1 最適な翻訳行列の計算法

逐次的に T を良くすることを考える。

$$F(T) = |T^t AT - B| \tag{7}$$

と置く。求めたいのは、

$$dF = F(T+dT) - F(T) < 0 \tag{8}$$

となる様な微小な dT である。T の成分を t_{ij}、dT の成分を dt_{ij}、などと書くと、全微分の公式より

$$dF = \sum_{i,j} \frac{\partial F}{\partial t_{ij}} dt_{ij} \tag{9}$$

であるから、適当な微小な変分の長さを $ds > 0$ として、

$$dt_{ij} = -\frac{\partial F}{\partial t_{ij}} ds \tag{10}$$

としておけば、

$$dF = -\sum_{i,j} (\frac{\partial F}{\partial t_{ij}})^2 ds < 0 \tag{11}$$

となって式 (8) を満たしつつ T を逐次的に更新することができる。したがって、dT を求めるには $F(T)$ を T の各成分 t_{pq} で偏微分すれば良い。$F(T)$ を書き下すと、

$$F(T) = \sum_{u,v} (\sum_{i,j} t_{iu} a_{ij} t_{jv} - b_{uv})^2$$

である。t_{pq} で偏微分することを考えると、それが現れるのは、$u = q$ または $v = q$ の時であるから、それらの項のみを u, v に関して対称性を考慮して特に明記すると、

$$F(T) = 2\sum_{v} (\sum_{i,j} t_{iq} a_{ij} t_{jv} - b_{qv})^2 + \text{その他の項}$$

となる。尚、B は対称行列であるから、$b_{qu} = b_{uq}$ である。t_{pq} が出てくるのは $i = p$ の時のみであることをふまえて t_{pq} に関して偏微分を行なうと、

$$\frac{\partial F}{\partial t_{pq}} = 4\sum_{v} ((\sum_{j} a_{pj} t_{jv})(\sum_{i,j} t_{iq} a_{ij} t_{jv} - b_{qv})) \tag{12}$$

である[3]。A が対称行列であることをふまえてこれを行列表現で書き下すと、

$$dT = 4AT(T^t AT - B)$$

となる。すなわち、

$$T_{n+1} = T_n - AT(T^t AT - B)ds \tag{13}$$

なる漸化式によって、逐次的に T を計算できる。$dT = 0$ となる点が最良の T であり、式 (3) が成り立つ。

実際は T の行和は 1.0 という制約があるため、以上の議論は Lagrange の未定乗数法を用いて T_{n+1} の逐次計算を定める必要がある。それによれば、行ごとに、T_n の行和から 1.0 を引いた値を T_{n+1} の行の各値に足しておけば、T の行和が 1.0 となるように収束する。

L_A の語数を N_A, L_B の語数を N_B とおけば、$F(T)$ は $N_A \times N_B$ 個の変数を持つ。T を行列ではなく、$N_A \times N_B$ 次元のベクトルと見なすと、grad $F(T)$ は $F(T)$ を増やす方向の法線ベクトルである。式 (9) に依れば、偏微分係数と同じ符号を持つように dT を定めれば dF は必ず小さくなるが、法線ベクトルの逆方向に T を変化させていけば、$F(T)$ は最も急速に最小化できる。したがって、

$$dT = \text{grad } F(T)$$

と置くという解釈ができる。

4.2 本手法の問題点

本稿で提案する手法を使って、ある語に対する正しい訳語を選ぶためには、L_A においてその語と共起関係にある語の訳語の集合が、L_B においても類似した共起関係を持つ必要がある。見方を変えると、類似した共起関係を持つ語の集合が L_B において複数存在する場合には、正しい訳語関係を定めるための橋渡しとなる情報が必要である。次章で示す実験では、電子辞書から得られた訳語対応を、この「橋渡しとなる情報」として用いている。

[3] $v = q$ かつ $u = q$ の場合を考慮していないのではなく、$v \neq q, u \neq q$, $v = q$ かつ $u = q$ の三つの場合に分けて考えると、結局このような形にまとまる。

上の事実は対訳コーパスを用いないことの代償として起こっている。対訳コーパスからの訳語関係の抽出は、T_0 が任意である点においては本手法よりも勝っている。これに対し本稿では、対訳コーパスだけがあって辞書がまったくない状態ではなく、対訳でないコーパスと原始的な辞書から訳語関係を抽出したいという状況下の問題解法を提示しているわけである。

4.3 過去の研究との関連

対訳コーパスを用いて対応頻度を取り出す処理は T を対訳コーパスから直接計算する手法である。しかし対訳コーパスの量の制限や、語レベルの対応付けは精度の限界から、訳語関係の抽出には以下の様な手法も提案されている。

1. 対訳コーパスは対応する文や語が必ずしも存在するとは限らないという、通常雑音を含んだものであるが、それを前提に名詞の二言語語彙情報を抽出する [Fun95]。まず統計情報から高頻度で確実に対応する語を抽出し、次にそれとの共起情報を用いて低頻度対応を抽出するという、二段階処理を行なう。

2. 電子辞書によって、結果を補正する [Uts94]。統計情報から、文対応、語対応をまず抽出する。対応の中に電子辞書上の対応があれば、それを電子辞書にない対応より重視するように補正する。補正には、電子辞書にある語とない語の L_B 内での共起を用いる。

いづれも、対応付けの限界を単言語内情報で補正してより豊富な対応頻度情報を抽出する方向であり、本研究とはその点で関係がある。

5 簡単な実験

英語に対し、日本語の訳語を選ぶ実験により、曖昧性の解消度を測った。30M の Wall Street Journal と 33M の朝日新聞の政治経済面の記事を形態素解析器[4]にそれぞれかけて、名詞、動詞、形容詞、副詞を取り出した。日本語は注目する語の前後各 5 語 (計 11 語) の窓を設けて、3 語以上と共起した語 (12871 語) から B を作成した。

5.1 局所文脈に依存した曖昧性の解消の検証

T の初期値は英和和英辞典 [Koi90][Ich90] より抽出したものを用いる。

[4]pc-kimmo と JUMAN を用いた。

表 1: doctor の例

	距離 pat	距離 squ
医者	3904	0.00576142
先生	3906	0.00576346
博士	3906	0.00576346
医師	3880	0.00553252

doctor と operation, hospital, medicine, nurse, patient, bed が共起したと仮定して 3.2 節で説明した計算を行なった。A の大きさは 7 で、対角成分が 0.0、他の成分はすべて同じ値として作成する (すなわち、これらの語はすべて自分以外の他の 6 語と共起したものとして A を作成した)。

T^tAT は 12871^2 の大きさとなることをふまえると、各訳語候補が正しいと仮定して T^tAT をそれぞれ計算するとそれらの結果は互いにほとんど等しいがほんの一部だけが訳語候補によって異なる。したがって、訳語候補によって計算結果に影響を与えるような行列の一部分のみ予め抽出し、$|T^tAT - B|$ を計算する。このような部分とは、上の 7 語から出発して共起関係、訳語関係をたどることのできるグラフである。

結果を表 1に示す。**医師**の方が**医者**よりも距離が小さいが、それは朝日新聞では**医師**が一般的であって、**医者**はほとんど用いられないためである。そのために、**医者**は訳語としては不適当となる。

次に、より大規模な実験を行なって曖昧性の解消率を算出する。英語の語は共起頻度 50 語以上の語をコーパスより抽出した。コーパスで共起頻度 50 以上の語が 11 語連なって生起する部分をランダムに取りだし、中央の語の曖昧性の解消を試みる。(すなわち前後 5 語を共起語とする)。ただし、実際に計算を行なう語には以下のような条件を付けた。

- 前後の文脈から、意味が確実に主観的に判断できるもの (すなわち、正解を判断できるもの。)
- 訳語候補が複数あって正解が訳語候補の一部分に絞られるもの。

たとえば、compare に対して、**比較する, 比べる**と**例える**の意味は異なる。前語の文脈を見てそれが、**比較する, 比べる**となるか、あるいは**例える**となるかを判断できる場合に $|T^tAT - B|$ を計算する。上の条件には、family income の income の訳

表 2: 曖昧性解消の実態

品詞	距離	正	誤	非解消
名詞	*pat*	81	29	17
	squ	85	27	15
動詞	*pat*	20	12	6
	squ	22	11	5
形容詞	*pat*	12	5	5
	squ	13	4	5
副詞	*pat*	8	4	1
	squ	8	4	1

語候補として、**稼ぎ**, **上がり**, **所得**が挙げられたような場合も含む。なぜなら、日本の新聞経済面では**上がり**などの口語が記述されることはまれであるので、**所得**が選択されることが妥当と考えられるからである。

距離が最も小さいものを訳語として、その妥当性を主観的に判断する。訳語に順序が付かなかった場合は曖昧性が解消されなかったものとする。$|T^tAT - B|$ を計算した語が総数 200 語となるまで計算した。結果を表 2に示す。

全体的な傾向としては、以下の点が挙げられる。

- 日本語コーパスに従った訳語が選ばれる。
  - worker に対して**労働者**よりも**従業員**
  - believe に対して**信じる**よりも**思う**
- 意味が明確に二分する語は正解率が高い。
  - bank に対して**土手**よりも**銀行**
  - patient に対して**忍耐強い**よりも**患者**
- 距離によって選ばれる候補が異なることがある。概して二乗和の方が正解率が高い。
- 名詞の正解率に比べると、動詞、形容詞は低い。

非解消の語は問題の語の周りの何語 (窓の大きさ) を共起語として用いるかに依存する。ここでは前後 5 語づつとしたが、窓が極端に小さい場合は非解消率は高く、大きくするにつれて一定値に近付くであろう。

5.2 大域的な翻訳行列の最適化計算

本稿では doctor に関して手法の特徴を定性的に解析する。doctor と以下の様な共起が起こったと仮定する。

- 実際に日本語のコーパスで共起した訳語 (17 語)
- **医師**の意味を強調するように故意に付け加えた語 (13 語)

doctor と上の 30 語とのみ共起し、他の語どうしは共起しないものとして A を作成する。T には doctor とその共起語 30 語の電子辞書に記載されていた訳語すべてと、さらに doctor に対して以下のような不適当な訳語候補をつけ加える。

情報 頻繁にどの日本語とも共起する語

乳母 **医師**と関係のある語

研究 **博士**と関係のある語

挨拶 まったく関係のない語

なお、電子辞書に記載されていた doctor の訳語候補はもともとは**医者**, **医師**, **先生**, **博士**であった。すなわち、doctor は 8 語の訳語候補を初期的に持つ。B は 3 語以上の共起頻度を示した 12871 語のうち、T に現れたすべての日本語 (154 語) とする。

T を逐次的に計算すると、**医師**、**情報**のみ訳語として生き残り、残りは全て翻訳確率が減る傾向を示した。各訳語候補の動向に関して次のような原因があった。

情報 **医師**の共起関係と、**情報**の共起関係はこの場合似ている。**情報**は共起する語が多いので、**医師**の共起する日本語と**情報**は共起する率が高い。すなわち、本実験においては**情報**は**医師**とほぼ同じ共起傾向を示し、正しい訳語と類似する共起傾向を示す訳語候補は落ちない。

医者 朝日新聞では**医者**は頻出しないため、共起情報も乏しく、たとえば**看護婦**などとも共起しない。このような候補は落ちる。

挨拶 B における共起情報がない。この場合は偏分係数は $-AT(T^tAT)$ となって、大幅に翻訳確率が減る。

doctor の他の訳語は**医者**と同様に、**医師**が共起する日本語と共起しないために落ちる。**医者**が訳語から落ちることは本手法の弱点ではない。なぜなら、A, B を作るもととなったコーパスに特殊化された T を作るのが提案する手法であるので、A

に医者があまり現れない状況では、これが一般的な訳語としてはいかに正しくとも訳語から落ちることが望ましいからである。

副産物として、doctor以外の訳語で、nurseと**看護婦**などの訳語関係は翻訳確率は増加の傾向にある。それは、nurseはdoctorとのみ共起したが、その訳語の**医師**と**看護婦**が共起するからである。

6 結論と今後の計画

本稿では対訳でない独立のコーパスを用いて局所あるいは大域文脈で訳語関係を抽出する手法を提案した。「第一言語で共起する二つの語の訳語は第二言語でも共起する」という仮定に基づいて、第一言語の共起情報を第二言語に翻訳することを確率行列の枠組で形式化した。翻訳した第一言語の共起情報と第二言語の元の共起情報の距離が小さくなるように翻訳行列を最適化することが訳語関係の曖昧性を解消することに等しいというのが提案する手法の要である。

文脈に基づいて訳語の選択を行なうという曖昧性の解消の問題をこの枠組で行なえることを示した。さらに、大域的に訳語関係を抽出するために、行列を逐次的に最適化する原理を示した。近年さかんに行なわれている対訳コーパスからの抽出とは異なり、本研究は非対訳コーパスからの訳語関係の抽出である。しかし、その代償として、正誤含めて訳語候補がすでに存在している状態から必要なものをとり出すという問題向きであることを示した。実際に簡単な定性的な実験を行なった。

今後の計画は大規模な翻訳行列の最適化をする実験を行う。これには巨大な行列計算を数値誤差を防いで収束させることが課題のひとつである。その際、最適な翻訳行列への収束速度も問題となるであろう。また、曖昧性の解消の精度が動詞や形容詞で低いので、その精度を上げるための手法を考えたい。

参考文献

[Fun95] Fung, P. (1995). A Pattern Matching Method for Finding Noun and Proper Noun Translations from Noisy Parallel Corpora. *Proceedings of ACL'95.*

[Rap95] Rapp, R. (1995). Identifying Word Translations in Non-Parallel Texts. *Proceedings of ACL '95.*

[Dag94] Dagan, I. and Alon, I. (1994). Word Sense Disambiguation Using a Second Language Monolingual Corpus, *Computational Linguistics*, vol. 20, pp. 563–596.

[Uts94] Utsuro, T. et al. (1994). Bilingual Text Matching using Bilingual Dictionary and Statistics. *Proceedings of the International Conference for Computational Linguistics '94,* pp.1076–1082.

[Bro93] Brown, P. et al. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, vol. 19(2), pp. 261–311.

[Che93] Chen, S. F. (1993). Aligning Sentences in Bilingual Corpora using Lexical Information, *Proceedings oc the 31th Annual Meeting of ACL,* pp. 9–16.

[Kay93] Kay, M. and Roscheisen, M. (1993). Text Translation Alignment, *Computational Linguistics*, vol. 19(1), pp.121–142.

[Bro91] Brown, P. et al. (1991). Word-sense Disambiguation Using Statistical Methods. *Proceedings of ACL'91.*

[Dag91] Dagan, I. et al. (1991). Two Languages Are More Informative Than One. *Proceedings of ACL'91.*

[Ich90] Ichikawa, S. (1990). *New Japanese-English Dictionary,* Kenkyuusha.

[Koi90] Koine, Y. (1990). *New English-Japanese Dictionary,* Kenkyuusha.

[Tok90] Tokunaga, T. and Tanaka, H. (1990). The Automatic Extraction of Conceptual Items from Bilingual Dictionaries. *PRICAI '90.*

[Chu89] Church,W. and Hanks,P. (1989). Word Association Norms, Mutual Information, and Lexicography. *In Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*, pp. 76–83, Vancouver, Canada.